AD_____

Award Number: DAMD17-00-1-0417

TITLE: Environmental Exposures at Birth and at Menarche and Risk of Breast Cancer

PRINCIPAL INVESTIGATOR: Jo L. Freudenheim, Ph.D.

CONTRACTING ORGANIZATION: State University of New York
Amherst, New York 14228-2567

REPORT DATE: June 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2001	Annual (1 June 2000 - 31 May 2001)

4. TITLE AND SUBTITLE
Environmental Exposures at Birth and at Menarche and Risk of Breast Cancer

5. FUNDING NUMBERS
DAMD17-00-1-0417

6. AUTHOR(S)
Jo L. Freudenheim, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
State University of New York
Amherst, New York 14228-2567
email: Jfreuden@Buffalo.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This is a population-based study that examines location of residence during the period between birth and menarche in relation to proximity to industrial sites, gasoline stations, toxic waste sites and heavily trafficked roadways as risk factors for subsequent breast cancer. We also will examine estimated exposure to benzene and PAHs as risk factors and evaluate genetic susceptibility in relation to these exposures and breast cancer. To date addresses for approximately 2,682 participants have been entered, and 200 participants' addresses have been geocoded using a Geographic Information System (GIS). 1,740 blood samples were sent for DNA extraction and genotyping. A validation study for geocoding was conducted to choose an efficient mapping tool, to determine accuracy of matched addresses, and to determine reasons for address matching failures. Based on the validation study, GDT/Dynamap will be used to geocode addresses in this study; ZP4 software and Polk Directories will be used to clean address data and to find missing residential information. We are just beginning to examine possible sources for the exposure assessment including aerial photographs of the study area, water sources, industrial directories, and the EPA hazardous waste site list.

14. SUBJECT TERMS
Breast Cancer, Cancer Etiology, Carcinogenesis, Environment Geography, Molecular Epidemiology, Gene-Environment

15. NUMBER OF PAGES
20

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION

In this population-based study we are examining environmental exposures experienced at birth and at menarche as risk factors for breast cancer. We will examine location of residence during these potentially sensitive time periods in relation to proximity to industrial sites, gasoline stations, toxic waste sites and heavily trafficked roadways as risk factors for subsequent disease. Complete residential histories were obtained from all participants during our existing case-control study. This study includes women, age 35-79 with incident, primary, histologically confirmed breast cancer living in Erie or Niagara counties. Controls are frequency matched to cases on age, race and county of residence. The residence at the time of birth and at menarche, as well as the potential exposure sites will be geocoded into GIS. The primary objectives of this study are: 1.)To investigate distance from steel mills, chemical factories, gasoline stations, toxic waste sites, other industrial sites and major roadways of the residence of cases and controls at the time of birth and at menarche as risk factors for pre- and postmenopausal breast cancer. 2.)To examine estimated exposure to benzene and to PAHs as risk factors for pre- and postmenopausal breast cancer. 3.)To evaluate genetic susceptibility in relation to these exposures and breast cancer risk by examining genetic variability in metabolism by NQ01, GST M1-1, GST P1-1 and CYP 1A1. Potential confounding factors will also be assessed. These include age, education, income, family history of breast cancer, Quetelet index, body fat distribution, having been breastfed, age at menarche, age at menopause, pregnancy history, lactation and contraceptive history, menstrual cycle length, birth weight, smoking and passive smoke exposure history, and diet and occupational history. There are no major results at this time. However, a validation study was conducted, comparing TIGER and GDT/Dynamap as mapping tool to geocode addresses. Accuracy of these geocoding processes was assessed and recommendations were made to improve geocoding matching rate and accuracy.

BODY OF REPORT

Task 1: Investigate distance from steel mills, chemical factories, gasoline stations, toxic waste sites and other industrial sites of the residence of cases and controls at the time of birth and at menarche as risk factor for pre- and postmenopausal breast cancer.

The following sections pertain to task 1 of the Statement of Work. Specifically, in order to assess residential distance from steel mills, chemical factories, gasoline stations, toxic waste sites, and other industrial facilities, addresses of cases and controls must be spatially located in relation to these sites. We are using Geographic Information System (GIS) to locate (geocode) addresses of cases, controls, and point sources of environmental exposure. Since this distance will be the primary basis for estimating historical exposure we examined several street databases for Erie and Niagara counties. This progress report begins with a brief discussion of geocoding, followed by a description of a geocoding validation study and concludes with a discussion of Polk Directory searches for missing addresses information.

Discussion of Geocoding Process:

Address geocoding in ArcView is a process that creates a theme based on the address data in a tabular form (event theme) and a reference feature theme (street map) to add point locations defined by the street address to the map. It consists of two major steps: 1) *Make a theme matchable*, 2) *Geocode addresses* (Add address events and locate an address).

When geocoding tabular data containing addresses, ArcView reads the addresses, finds where they are located on your map, and creates a new theme containing a point for each address it was able to find. In order to use a reference theme in geocoding, it must have a *geocoding index*, an index that ArcView builds to speed up the process of finding addresses in the theme.

The most frequently used address format is the "*US street address with zone*" format. The street address is contained in one field. Another field contains the Zip code, which ArcView uses as the zone identifier to resolve situations where there is more than one street with the same name in a given area. A field containing the city name can be used as the zone identifier instead of the Zip code. The fields containing the street address and the zone identifier can have any name. Address data typically also has a field containing the state name or abbreviation.

ArcView assumes that the geocodable theme's attribute table is in a standard format. Geocoding match scores may be unexpectedly low if the attribute table does not conform to the standard. The standard format requires each address component to be stored in separate fields and that each address component uses abbreviation standards. Depending on the geocoding style being used, the table should include some of the following fields, Predirection, Pretype, Street Name, Post Direction, Post Type, and Zone. Abbreviation

standards include: "AVE" for Avenue, "ST" for Street, "RD" for Road, "PK" for Pike, "E" for East, etc. The Standardization rules are also applied for the address database.

One way to create a standardized address table is to use ArcView programming language "Avenue." The script will parse out information in the address field to the appropriate component fields using the predefined abbreviation standards. Although ArcView supports a wide range of abbreviations, we can modify the "US_ADDR.cls" or "STNAME.cls" file to add more abbreviations. The .cls files are used to classify particular words and abbreviations.

Address matching depends not only the quality of the reference theme, but also the tabular data to be mapped. The first trial gave a 60% matching rate, but the address data contains many inaccurate and non-standardized formats. There are several ways to improve the quality of both datasets of course. Subsequent review of 500 participants' address data revealed several common problems:

1. Problems of Reference Theme (Street Map):

The most commonly used Census Bureau's TIGER/Line files are called Centerline files because they depict the center of the street by line segments. Details of curbs, alleys, and cul de sacs are not depicted in Centerline files. TIGER/Line files are designed to show only the relative position of geographic elements. Thus, they don't have high levels of positional accuracy, thus may result in inaccurate location of address on the street. An incomplete or an out of date street map may result in unmatched or out of range addresses. The most common problems for street maps are range errors, i.e., the address identified is out of address range of the street map.

2. Problems of Address Database:

Blank spaces before or between addresses
Street name errors such as misspelled or ambiguous street names;
Example: 333 1 street instead of 333 1st st.

Unusual or non-standard abbreviation;
Example: 222-11 elm aven. instead of 222 elm ave.

Wrong or out of date zip code (boundary change)
Business address such as school or hospital address
Unnecessary words like "near"
Inconsistent street types
Apartment letters or numbers
Common name for the known location such as freeway or highway.
Example: interstate 290 instead of I 290, route 39 instead of US hwy 39, rd 1 instead of an official name

Other Problems:

Multiple matching: Several locations of address in different parts of the same region. Some street names preceded by N or S should be recognized as part of the street name, rather than a directional. ArcView will recognize North as directions by default.

Intersection address: ArcView geocode intersection address given as streets separated by an ampersand, but street intersection address given by other symbols may cause problems.

Some of the above problems were overcome with several methods. For example, by constructing a type of alias table for common location and misspellings problems or using "find-and-replace" with Microsoft Word or more complex queries with Microsoft Access. However, most of the major problems were overcome by:

1. Purchasing a more accurate street basemap:

GDT/Dynamap is a commonly used street map for geocoding purposes. We compared the accuracy of GDT and TIGER with Aerial Photo for the 14221 zip code area.

The test results for one zip code, 14221, using GDT: showed a 14% increase of matching rate. For zip code 14221, GDT had 300 more street segments than TIGER. The street field was separated as street name, type, prefix, suffix while TIGER doesn't have the prefix and suffix fields. Unlike TIGER, GDT/Dynamap has more up-to-date information than TIGER and a better positional accuracy. GDT/Dynamap: estimated price for 120 zip codes in Erie and Niagara County - $10 per each zip code.

2. Using the stand-alone address cleaner ZP4:

To standardize the addresses to be matched with USPS certified addresses, standardization software ZP4 can correct and update out of date or incorrect information. Depending on original address data, matching rates can be improved by 20%.

3. Data Entry
Separate each column and follow the standard format for geocoding purpose

<u>Geocoding Validation Study:</u>

Geographic data is of increasing importance in epidemiologic research. (2,4,5,6). Topographically Integrated Geographic Encoding and Referencing System (TIGER) and GDT/Dynamap database are all datasets commonly used to geocode addresses. While TIGER is free, it costs approximately $10 per zip code area to purchase the GDT/Dynamap database. A database that permits complete and precise geocoding is essential. We conducted the following validation study to compare the TIGER and GDT/Dynamap file as reference themes for mapping addresses to actual locations. The objectives were to:

1. Compare matching rates for Buffalo and the region outside Buffalo.
2. Determine the accuracy of TIGER and GDT/Dynamap files locations for matched addresses.
3. Determine reasons for address matching failures.
4. Compare TIGER to GDT/Dynamap database.

A sub-set of 20 participants (89 addresses) was randomly selected for this validation study. All participants were women, aged 35-79, residents of Erie or Niagara County in western New York State. Self-reported lifetime residential histories were collected. Of the total addresses reported, only the 89 addresses in Erie and Niagara County were used in this study.

To examine if the TIGER and the GDT/Dynamap reference themes were more accurate in the City of Buffalo as compared to outlying areas, the total 89 addresses were grouped into those addresses located in City of Buffalo, and those located outside the city. Locations outside Buffalo in this study included Alden, Amherst, Blasdell, Boston, Cheektowaga, Clarence, Depew, E Amherst, Grand Island, Hamburg, Kenmore, Lewiston, Lockport, Middleport, North Collins, North Tonawanda, Niagara Falls, Orchard Park, Ransomville, Snyder, Springville, Tonawanda, West Seneca and Williamsville.

Addresses were geocoded using ArcView 3.2 using TIGER as the reference theme. Addresses were located according to the street number, street name, and zip code. After performing the initial geocoding, addresses that matched were compared to historical Sanborn maps for the City of Buffalo or by driving to the address and comparing its actual location to the location placed by the TIGER map.

The matching rates for the TIGER file are depicted in Table 1 in the Appendix. Among the 89 addresses, 55 (62%) were matched using the TIGER file and 34 (38%) addresses where the TIGER file failed to match. There were 27 addresses from the City of Buffalo, and 62 addresses from outlying areas. In the City of Buffalo, the automated matching rate was 85% (23/27) and outside Buffalo, the matching rate was 52% (32/62), much lower matching rate.

Of the 23 addresses in the City of Buffalo that matched, TIGER accurately located 21 (91%) addresses when compared to the historical Sanborn Maps. The other 2 addresses (9%) addresses were not accurately located. For one address the distance from where TIGER located the address and where it was located in the Sanborn Map was larger than 0.1 mile. In fact, it was approximately 0.20 miles apart. For the other inaccurate address, TIGER chose the wrong location when there were several matching candidate locations available. This happened when study participants did not provide the zip codes for their addresses. Interestingly, 5 of the matched addresses were for schools, associations, or churches. Historical Polk Directories were checked to make sure the participants accurately recalled their residential addresses.

Of the 32 addresses outside of Buffalo, TIGER accurately located 25 (78%) addresses when compared to drive-bys. TIGER failed to accurately locate the 7 (22%) remaining

addresses. Distance errors occurred in 6 (19%) addresses with an average distant error 0.24 miles. For the one remaining inaccurate address, TIGER selected the wrong location when there were several candidate locations available.

Thirty-four addresses failed to match using the TIGER as a reference theme. The majority of these failures (88%) occurred in those addresses lying outside of the City of Buffalo. These failures occurred because (1) the reported address was incomplete (i.e., missing street number), (2) the zip code was wrong, or (3) TIGER error (street number range error). TIGER errors usually refer to an errors in street number designations. Typically, TIGER files have beginning street numbers and ending street numbers for any given street section. For example Pleasant Ave between Lake Ave and Sunset Blvd. may range between 23 and 34. In the TIGER file, however, the ranges may be listed from 43 to 99 as shown in Figure 1 in the Appendix. These out of range errors occurred in 29% (10/34) un-matched addresses.

For the incomplete reported addresses, Polk City Directories are used to find a complete address. For addresses with an erroneous zip code, ZP4 software automatically updates the zip codes based on the town name.

The validation was duplicated using GDT/Dynamap as the reference theme and Table 1 shows that the matching rate was high when GDT/Dynamap was used. Except for the incomplete reported addresses, all other addresses were matched both in the City of Buffalo and outlying areas. Further, the identical matching rates are also very high in both the City of Buffalo (100%) and outlying area (93%). These rates are all higher than those using TIGER as the reference theme.

Table 2 in the Appendix shows the differences in matching rates between the TIGER file and GDT/Dynamap reference themes. Overall, the GDT/Dynamap matching rate was more efficient (78% vs. 62%). We also found that GDT/Dynamap, not only identically matched all the addresses that were identically matched by TIGER, but that 4 other addresses with previous distance errors in TIGER were identically matched. Although, there were still 3 (7%) addresses with distance errors outside Buffalo, the average amount of error was smaller with GDT/Dynamap than with TIGER. Outside Buffalo, the average distance error was 0.16 miles with GDT/Dynamap and 0.24 miles with TIGER. Lastly, 1 address in the City of Buffalo and 9 addresses in outlying areas that were previously not matched by TIGER because of TIGER errors, were identically matched by GDT/Dynamap (some were identically matched after using ZP4 to clean the address data).

We found that both TIGER and GDT/Dynamap were more efficient in the City of Buffalo than in the outlying areas of Buffalo. These results were similar to a previous report carried out in North Carolina where automated address-matching rates were as low as 20% in very rural counties, and as high as 98% in large urbanized counties. (5).

The North Carolina study used the TIMS (Transportation Information Management System) spatial dataset. TIMS is used for school bus routing and the assignment of bus

stops. But the address-matching rate was also very low in rural area. Further, commercially available spatial datasets are expensive and may be only good for gecoding some regions, and not the others depending on several factors. Furthermore, we found that GDT/Dynamap, a commonly used commercial dataset for geocoding, has a much higher matching rate than TIGER, and was more accurate than the TIGER file.

There are some significant differences between a study using historical residential data and one that is using current residential data. In this study, many addresses have missing information and require extensive resources to find historical data. Historical address information was not complete for 19 addresses. If we exclude these incomplete addresses, the address matching rate for TIGER was 79%(55/70), and for GDT/Dynamap was 100%(70/70). Furthermore, there are no historical reference themes. It is possible that street names have change, neighborhoods which once consisted of houses and apartments may now be shopping malls or parking lots, streets may have been re-numbered and historical street numbers may no-longer coincide with current locations.

Polk Directory Search:

A Polk Directory search is done to find missing addresses, addresses with missing street numbers, and to verify the address reported by the subject. The primary Polk search begins with sorting the residential data file by address to obtain a list of addresses missing street numbers or with blank street names. A new file is created with only the missing entries. Once the missing entries have been identified, the subjects' current name and maiden name are retrieved. This is necessary to identify candidate addresses listed by name in the Polk directories. A second data file is created with subjects' names and addresses, city, state, zip code, year moved in, year moved out, and date of birth of the subject. This list provides all the basic information required to conduct a Polk Directory search.

Polk Directories were compiled for numerous cities and geographic localities across the United States. These directories for the City of Buffalo, and North East Suburbs, South East Suburbs, Hamburg, and the Tonawanda's are all readily available at the Eire County Library. In addition, some volumes for these locations are kept in the Dept. of Social and Preventive Medicine. Each volume was designated for a specific time period and for a specific location. For example, one volume contains the names of individuals residing in the City of Buffalo for 1957. Another volume will have residents for Williamsville, Cheektowaga, Amherst, Sloan, and Eggertsville for 1978 in the North East Suburbs volume.

The Polk Directories have two sections. One section lists individuals by surname and first name. These entries not only include the address, but also supply some information about the individual's occupation, and spouse's name. This supplementary information is very useful for finding the correct address for individual subjects. The second section of the Polk lists residents by the street name in alphabetical order and lists the house numbers on that street. It will also indicate vacant addresses and will specify the apartment number (letter) for residents of apartments.

There are several important caveats to the Polk directories that must be considered. First, the Polk directories are not comprehensive. All households are not accounted for. Secondly, the compiling methods resulted in some important errors. In some instances new residences were not updated and residents can be listed at old addresses. Furthermore, spelling errors in names can lead an inability to locate an individual. Third, people do not move into new residences on January 1st and depending on the time of year the Polk directory updated the listing and when the person moved, this may not be reflected in the new volume. Lastly, the study subject's first name, when she was a child, will not be listed. Therefore, we must rely on her maiden surname as an indicator of address. This is sometimes problematic when there are more than one of the same surnames listed for a street.

To date, this type of search has been the most comprehensive method identified for obtaining historical address information. For the GIS-Breast Cancer Study we developed a disposition sheet to facilitate the Polk directory search. This sheet (a copy is in the Appendix) aids in the search by tracking whether the maiden and/or surname was searched and for which years.

The procedure is rather simple, but time intensive. Basically, the Polk Directories are searched for the subject's name based on the year they report living at an address. The precision of the search can be improved by taking year of birth into account. For example, if the subject was 5 years old in 1965 and resided on Lake Ave, it is logical to search the Polk for the subject's maiden name. Because the subject is so young it really does not make sense to investigate the current name. However, in the event that the maiden name is not available, a search by the current name may be beneficial if the subject has never been married or retained her maiden name after marriage.

The disposition sheet allows for the recording of other important information, which may provide clues to the subject's address. As mentioned previously, the Polk sometimes included information about occupation. This can be useful to determine the correct address when there are several of the same surnames present or the street name is unknown. The Polk can be searched for a year where the subject's entire address is known. In this case the father's first name and occupation can be obtained and compared to the names listed for the unknown address.

Sometimes the subjects name is so common that there will 5-10 pages of that name to review. The street index is very helpful in these circumstances because that can be perused to find the name in question. However, this method should not substitute for finding the name in the name listing because that address may be for someone with the same name and not the subject. Furthermore, on rare occasions the name of interest is listed in one section, but not in the other.

We have compiled 5,239 addresses that are ready for geocoding. Of these 5,239 addresses, 1,565 had some missing information that is required for geocoding. Polk Directory searches were conducted for 12% (569/1565) of these addresses because there

either were no known dates of residence or the Polk Directory for the years/town when the subject resided at that particular address was not available. See Table 3 in the Appendix. The success rate for finding the missing address information in the Polk Directory was 34% (198/569).

Currently, we are conducting Polk searches for another 4,969 addresses of which 564 addresses have missing information that need Polk searches. This brings the running total ready for geocoding up to 10,208 addresses. We anticipate having over 15,000 total addresses to geocode covering the lifetime residencies of these breast cancer cases and controls.

Concern about the relatively low overall success rate for finding missing addresses has led to investigations of other possible strategies for finding address information. To improve the success rate, targeted searches in Town Clerk's Offices and of property records will be conducted to find subject's addresses at birth and near menarche only. In this way, addresses only pertinent to our hypotheses about early life exposure and future breast cancer will be explored, greatly reducing the time required to complete such a time intensive task.

Task 2: To examine estimated exposure to benzene and to PAHs as a risk factor for pre- and postmenopausal breast cancer, with control for the appropriate confounders.

We are beginning to investigate point sources of pollution and as of yet have not begun estimating exposure to benzene or polycyclic aromatic hydrocarbons.

Task 3: To evaluate genetic susceptibility in relation to these exposures and breast cancer risk by examining genetic variability in metabolism by NQ01, GST M1-1, GST P1-1 and CYP 1A1.

This task is currently in progress. We have met with Dr. Peter Shields to discuss the investigation of the above genotypes and to finalize protocols for collaboration with his laboratory (the Lombardi Cancer Center, Georgetown Medical Center). Blood clots, representing about 50% of the total, have been removed from the freezer and shipped on dry ice to Dr. Shields for analysis. To date we have sent 1,740 samples: 560 from breast cancer cases, 560 pairs of controls, matched to cases 2:1, and 60 randomly positioned blind duplicates.

Preparations are underway for a second and final shipment, similarly arranged, to be sent immediately following the termination of interviewing in the case-control study in which the samples are drawn. Drawing for that study is scheduled to end June 30, 2001.

We have found that it is also possible to extract DNA from collected urine and saliva samples. We will also extract DNA samples from urine or saliva from those participants who gave consent but were either unable or unwilling to provide a blood specimen.

KEY RESEARCH ACCOMPLISHMENTS

There are no key research accomplishments regarding the primary hypotheses. However, major steps towards testing our hypotheses are being completed.

- Addresses for 2,682 participants (13,487 individual addresses) were entered into database that permits geocoding of the addresses.
- Addresses for 200 participants (1,068 individual addresses) were geocoded with ArcView.
- 1,565 addresses were investigated using Polk Directories for missing street information. Strategies for targeted follow-up of missing address information were developed.
- 1,740 blood samples were shipped for DNA extraction and genotyping to Dr. Peter Shields (Lombardi Cancer Center, Georgetown University).
- A validation study of the geocoding methodology has been completed.

REPORTABLE OUTCOMES

Abstract for validation study for geocoding was accepted at the "Congress of Epidemiology 2001" in Toronto and a poster based on this abstract was presented on June 14, 2001.

Validation of TIGER (Topologically Integrated Geographic Encoding and Referencing system) to Geocode Addresses for Epidemiologic Research
Jing Nie, Matthew R. Bonner, Dominica Vito, Nicholas H Willett, Jo L. Freudenheim
Department of Social and Preventive Medicine, University at Buffalo.

Geographic data is of increasing importance in epidemiologic research. TIGER is one of the dataset commonly used to geocode locations. We used ArcView3.2/TIGER to geocode lifetime residential histories for 20 participants (89 addresses) from an ongoing case-control study. There were 55 addresses (62%) that matched by automatic geocoding. Addresses were divided into those in the city of Buffalo (27), and those in Buffalo suburbs (62). The matching rate was lower for suburbs than for the city, 52% vs. 85%. We validated the matched addresses in Buffalo by comparing locations generated by TIGER with historical Sanborn maps. Because historical maps were not available, we validated the matched suburban addresses by comparing locations generated by TIGER and by driving to the location. All addresses were within 0.6 miles. Only 16% were more than 0.1 miles for the suburbs, 4% for the city. The two major reasons for unmatched addresses were: 1) incomplete or inaccurate information for the reported address (75% of

unmatched addresses in Buffalo, 70% in the suburbs), which could be solved by using Polk directory and other sources to find the correct address before geocoding; and 2) inaccurate range of street numbers in TIGER (25% of unmatched addresses in Buffalo, 30% in the suburbs). Based on this study, using TIGER to do geocoding is accurate, although it may be less complete for suburban than for city populations and may require some manual inputs.

CONCLUSIONS

We are still in the process of geocoding residences and locating exposure sites. There are no conclusions regarding the major hypothesis to report at this report.

REFERENCES

1. Address Matching. http://www.uiowa.edu/~geog/health/address/address.html. Oct.03, 2000.
2. Betts KS. Mapping the environment. Environmental Health Perspectives. 105(6):594-596, 1997 June.
3. Getting to know Arcview GIS, the geographic information system (GIS) for everyone. Environmental systems research institute, Inc. 3rd edition, 1999. p26-1 – p26-24.
4. Howe HL. Geocoding NY State Cancer Registry. AJPH. 76(12):1459-1460, 1986 Dec.
5. Vine MF, Degnan D, Hanchette C. Geographic Information Systems: Their Use in Environmental Epidemiologic Research. Environmental Health Perspectives. 105(6):598-605, 1997 June.
6. White E. Geographic Studies of Pediatric Cancer near Hazardous Waste Sites. Archives of Environmental Health. 54(6): 390-397, 1999 Nov./Dec.

APPENDICES

Table 1. TIGER File and GDT/Dynamap Matching Rate and Accuracy for the City of Buffalo and Outlying Areas

	TIGER File	GDT/Dynamap
City of Buffalo (n=27)	# of Addresses (%)	# of Addresses
Matched		
Identical	21 (91)	25 (100)
Distance error	1 (4)	-
Wrong Address selected	1 (4)	-
Sub-total	23 (100)	25 (100)
Not-matched		
Incomplete Address	2 (50)	2 (100)
Wrong zip code	1 (25)	-
Inaccurate range of street #	1 (25)	-
Sub-total	4 (100)	2 (100)
Outlying Areas (n=62)		
Matched		
Identical	25 (78)	42 (93)
Distance error	6 (19)	3 (7)
Wrong Address selected	1 (3)	-
Sub-total	32 (100)	45 (100)
Not-matched		
Incomplete Address	17 (56)	17 (100)
Wrong zip code	4 (13)	-
Inaccurate range of street #	9 (30)	-
Sub-total	30 (100)	17
Overall Total	89	89

Figure 1. Inaccurate range of street number in TIGER

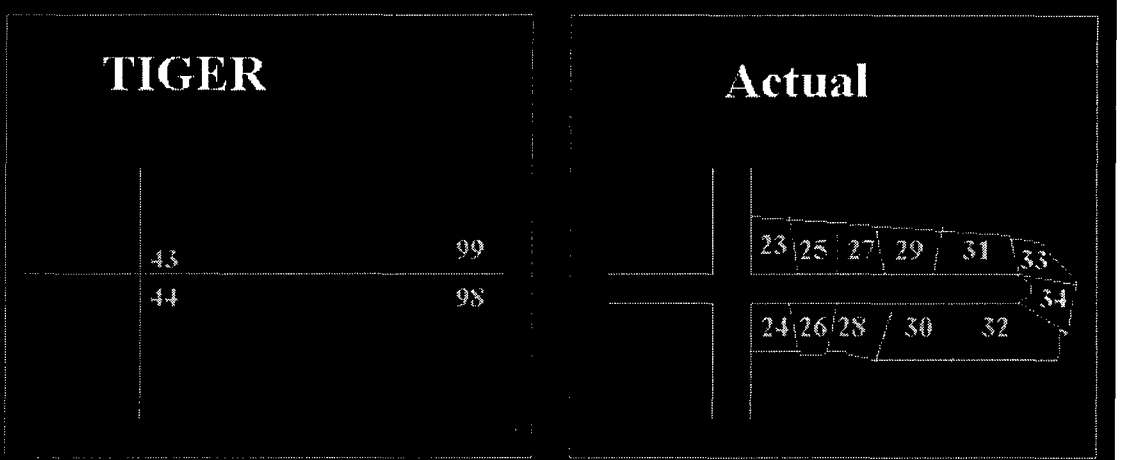

Table 2. City of Buffalo and Outlying Areas combined comparison of TIGER file and GDT/Dynamap reference themes.

	TIGER File	GDT/Dynamap
Matching Rate	62%	78%
Accuracy of Matched Addresses		
Identical	87%	96%
Distance error	13%	4%

Breast Cancer & GIS Study
Residential Geocoding Disposition Sheet

Respondent Number:

Re-Match Completed_____ DATE:_____

Address:
Street_____
City_____
Zip Code_____

Corrected Address:
Street_____
City_____
Zip Code_____

Reason for Matching Failure:

_____Missing Street Number
_____Misspelling of Address
_____Wrong Zip Code
_____Other:
_____TIGER Street Numbers Out of Range
_____Missing Zip Code
_____TIGER Error

Solution:

_____Yahoo Search
Results:

Completed Date:_____

_____Polk Search
Results:

Year Searched	By Current Name	By Maiden Name

Date Completed:_____

_____Other Search
Results:

Date Completed:_____

Table 3. Polk Directory Search Results (1,565 addresses with missing information)

Total addresses to date	5,239		
# Of addresses with missing information	1,565		
# Of addresses searched in Polk	569		
	# Found	# Not Found	# Pending (need maiden name)
City of Buffalo	76	77	78
Outlying Areas	122	125	91
Total	198	202	169
% Among searched (n=569)	34.8%	35.5%	29.7%
% Among # with missing address information (n=1,565)	12.7%	12.9%	10.8%